新闻媒体领域中文语义分析技术智能化、知识化之路的研究与探索

摘　要：媒体融合发展是一项复杂的系统工程，离不开技术系统的变革与创新。在新闻媒体领域数据爆炸，同时人工智能领域飞速发展的大背景下，本文针对国内新闻媒体领域中文文本语义分析过程中存在的诸多难题和现状，对中文文本语义分析在新华社业务系统中的智能化、知识化的探索之路进行阐述与展望。

关键词：中文语义分析；新闻媒体领域；智能分析；知识分析
中图分类号：TP391　　**文献标识码：**A
文章编号：1671-0134（2018）08-035-03　　**DOI：**10.19483/j.cnki.11-4653/n.2018.08.009

文 / 李泽魁　孙霏　陈珺

引言

在媒体格局、舆论生态、受众对象、传播技术都在发生深刻变化的今天，数据在新闻媒体转型发展过程中已成为全新的富矿。以新华社为例，一条新闻从生产源头的“采编发供”到用户读者端的传播与反馈，都离不开文本作为文学的载体和传播的媒介。这些蕴含着巨大潜力的文本大数据，合理、充分地挖掘其价值很有必要。

另一方面，伴随着自然语言处理技术的飞速发展，文本语义分析已经从 20 世纪基础的词典规则匹配、统计学概率计算的方法，渐渐转变为当前主流的机器学习、深度学习的智能分析算法。同时，分析对象与应用场景也越来越广泛，涵盖了包括新闻、评论、社交媒体等领域的各个方面。

党的十八大以来，以习近平同志为总书记的党中央高度重视传统媒体和新兴媒体融合发展。新华社作为媒体融合发展的排头兵、先行者，每天都需要对社内数万条稿件及海量的互联网文本进行实时准确的分析，中文语义分析作为基础技术，不可或缺。

1. 新闻媒体领域智能化的中文语义分析技术

1.1 结合新闻稿件特性的智能化词语切分

新华社日均有数以万条的稿件需要进行文本语义分析，而种类繁多的智能分析的背后，都离不开对文本进行词语的切分，即自然语言处理中的分词算法。众所周知，英文文本以空格切分单词，而中文文本需要根据语义切分词语，对连续字符按照语义规范进行重新组合，切分难度更大。针对新闻媒体领域的歧义识别与新词挖掘等中文语义分析难点，着力从三个方面对其进行智能化探索。

1.1.1 新闻媒体分词词库的自动化挖掘

实际应用的分词系统往往是多种算法的融合，但一般都依赖一套高精度的新闻媒体行业词库。为此，结合我社稿件文本特点，提出了基于共现词频过滤的新词发现、少量人工校验辅助的分词词库挖掘算法，一定程度上提升了分词准确率。

1.1.2 构建大而全的新闻媒体领域语料库

除了基于词库规则的分词算法，还有一种是基于统计机器学习的方法。这种方法依赖一定数量的“机器学习的教材”，即标注好正确切分结果的训练数据（语料）。为使分词模型更适合我社业务需求，我们收集了人民日报、国家语委、各大评测等高质量标注的训练数据集，充分利用新闻媒体领域的汉语组词的规律切分词组。

1.1.3 针对实体短语进行优化加强

作为国家通讯社，新华社从诞生起就在党中央的直接领导下开展工作，肩负党和人民赋予的神圣使命，发挥喉舌、耳目、智库和信息总汇作用。当然，稿件也以正确舆论导向与时代主旋律为主。为此，我们针对部分时事政策类的实体词组进行了大力优化，例如“一带一路”，“供给侧改革”等，提高了相关词组的切分能力。具体效果如图 1。

通用分词算法

建设 | “ | 一带 | 一路 | ” | ， | 是 | 以 | 习近平 | 同志 | 为 | 总书记 | 的 | 党中央 | 主动 | 应对 | 全球 | 形势 | 深刻 | 变化 | 、 | 统筹 | 国内 | 国际 | 两个 | 大局 | 作出 | 的 | 重大 | 战略决策 | 。

结合新闻稿件特性的分词算法优化

建设 | “ | 一带一路 | ” | ， | 是 | 以 | 习近平 | 同志 | 为 | 总书记 | 的 | 党中央 | 主动 | 应对 | 全球 | 形势 | 深刻 | 变化 | 、 | 统筹 | 国内 | 国际 | 两个大局 | 作出 | 的 | 重大 | 战略决策 | 。

图 1　结合新闻稿件特性的智能化词语切分效果对比图

1.2 基于知识属性的智能化主题分类

文字新闻报道是新华社的传统报道形式，也是核心报道形式之一。它及时、准确、权威地报道党和国家的方针政策及国内外时政、经济、军事、外交、文化等领域的重要新闻。为了更好地对我社文字稿件进行智能分析、智能检索与推荐，一套新闻稿件智能主题分类算法很有必要。

当前，新华社知识属性为多类、多级体系（13 种一

级知识属性分类、千余种多级指数属性分类）。结合这套知识属性，我们建立了多级主题分类体系（为了保证智能分类的准确性，最深可达二级分类，详见表 1），同时结合当前流行的深度神经网络算法，训练出一套可靠、高效的智能主题分类算法。

表 1　基于新华社知识属性的智能化主题分类体系举例

分类层级	类别举例
一级分类	政治、法律
	文化、艺术及娱乐
	……
二级分类	社会—社会福利、社会保障
	军事—武装力量及其活动
	……

1. 3 多个角度智能化情感分析

新华社在重大新闻报道上，除了要打赢新闻首发权抢夺战，同时也要兼顾热点事件的全方位、多维度的精准统计与分析，这样才可以始终保持舆论导向的正确性。

情感分析作为中文语义分析的一项基础任务，又称倾向性分析或意见挖掘。新闻领域的情感分析是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。

对热点事件新闻及评论进行情感分析，有助于对互联网舆论的全面监测与管理。在提升负面信息发现处置、情报预警和舆情导控能力的同时，又充分利用互联网数据服务于新闻生产全流程。为此，我们提出了从同一热点事件的不同角度进行深度情感挖掘的算法，各个话题的情感立场在界面中会一目了然地展现。如图 2 所示。

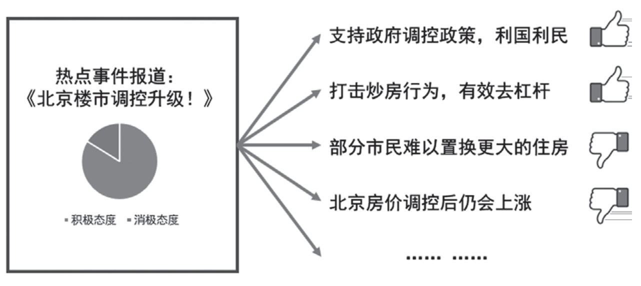


图 2　面向事件多个角度的智能化情感分析效果演示

1. 4 文本主旨的智能化自动摘要

自动文本摘要是利用智能化算法自动编写和生成摘要。面向新闻文本的自动摘要技术是解决当前我社大量稿件素材信息过载问题的一种辅助手段，有助于“采编发供”流程中各类用户更加快速、准确、全面地获取新闻文本信息。如何对这些新闻文本进行高效存储、信息检索与挖掘成为一个迫切需要解决的重要问题。

针对新闻领域智能化自动摘要的应用场景，结合新闻文本结构、句法及语义相关的知识特征，通过大量的迭代优化与试验，提出了面向新闻文本主旨的智能化自动摘要方法。

2. 新闻媒体领域知识化的中文语义分析技术

2. 1 结合新闻要素和特性的知识标签体系

众所周知，西方新闻界首先提出新闻要素的概念，即何时、何地、何人、何事、何故、如何。

为了使新闻文本要素与新闻知识标签抽取相衔接，让机器更加规范、智能地自动提取新闻标签，我们提出了新闻的标签体系，包括时间、地点、人物、概念、事件五类。其中，概念标签和事件标签的定义本文拟定如下：

概念标签：可概括为语义概念的文本词条实体。

事件标签：可表征事件的文本词条，直接引发事件的产生，是决定事件类别的关键特征。

其分类与举例详见表 2。

表 2　概念标签分类与举例

标签类别		标签举例
时间标签		“2017 年 4 月 1 日”“星期一”“上午 8 点半”等
地点标签		“北京”“西大街 97 号”“后海”等
人物标签		“特朗普”“爱新觉罗・溥仪”“小王”等
概念标签	通用知识库概念	“部门”“记者”“金融”等
	短语知识库概念	“党的群众路线”“供给侧改革”“伦勃朗效应”“高温补贴”等
	长尾知识库概念	“抗帕金森治疗”“名人婚纱设计师”“基础的水彩技巧”等
事件标签		“通报”“巡视”“经济增长”“军事合作”“召开会议”等

本文涉及的新闻体系结构图如 3：

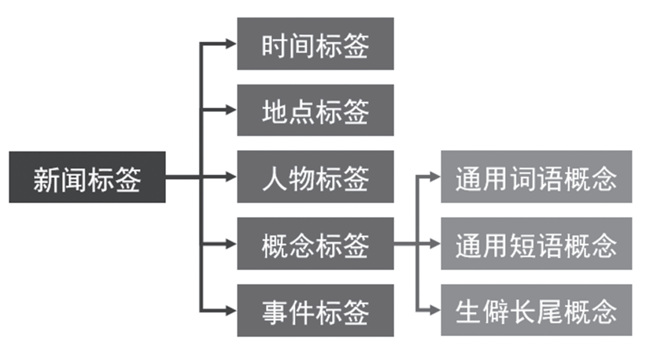


图 3　新闻体系结构图

2. 2 基于标签类别与权重的知识自动提取

面对铺天盖地的各类型新闻与素材数据，如何从中挖掘出真正有用的信息，是大数据应用的一道门槛。以我社稿件文本为例，在大量数据面前，本文首先提出了知识标签体系规范，再根据规范将稿件按时间、地点、人物、概念、事件等要素进行标注。具体算法分为基础中文语义智能分析、基于语义紧密度挖掘的短语合并、标签候选集的生成与过滤和依据语义关键度的排序输出

等步骤，如图 4 所示。

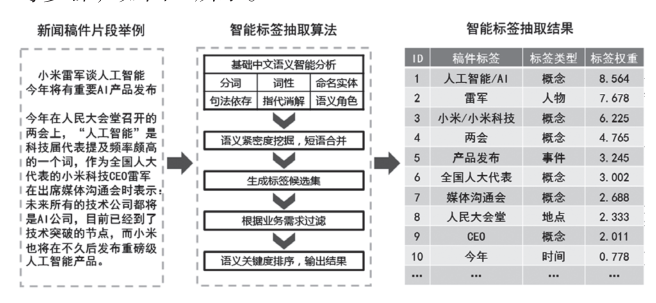


图 4　基于标签类别与权重的知识标签自动提取样例

伴随富标签体系的建立与智能抽取算法的设计，新华社现有稿件分类与检索存在的诸多问题将进一步缓解。同时，下一步我们会继续提升系统，以满足数字网络时代用户对稿件精细搜索、智能检索及个性化定制的需求，提高稿件存储和检索的高效性与准确率，深度挖掘稿件在不同领域的应用价值。

2. 3 面向业务系统的知识图谱初探

知识图谱作为知识工程的一个重要分支，以语义网络作为理论基础，并且结合了自然语言处理和知识表示和推理等优秀算法，在大数据的推动下受到了业界和学术界的广泛关注。

构建知识图谱的主要目的是获取大量有关联的、计算机可理解的知识网络。新华社建社之日起，八十余年的历史中，海量非结构化的稿件文本、半结构化的表格和网页以及生产系统的结构化数据中蕴含了大量待挖掘的新闻知识与关系（如图 5 所示），这部分资源犹如待开发的金矿，非常宝贵。

知识词条智能挖掘：
两会、人工智能、全国人大代表、人民大会堂
小米科技、CEO、AI……

知识词条详情展示：
两会：两会是历年召开的中华人民共和国全国人民代表大会和中国人民……

今年在人民大会堂召开的两会上，“人工智能”是科技届代表提及频率颇高的一个词，作为全国人大代表的小米科技CEO雷军在出席媒体沟通会时表示：未来所有的技术公司都将是AI公司，目前已经到了技术突破的节点，而小米也将在不久后发布重磅级人工智能产品。

知识共指智能连接：
1) 全国人大代表～小米科技CEO～雷军
2) 人工智能～AI
3) 小米科技～小米

知识词条智能推荐：
1) 两会：房产税、姚明、人民大会堂……
2) 人工智能：自动驾驶、深度学习……
3) 雷军：小米、乔布斯、饥饿营销……

图 5　面向新华社业务系统的知识图谱初探

知识图谱构建，包含了许多关键性技术。从较为基础的自然语言处理技术，对稿件文本进行较为精确的分词、实体提取、句法识别等工作，到进阶的实体关系识别、知识融合、实体链接和知识推理技术等。

鉴于垂直领域词典匮乏、知识人力标注成本高等现状，当前新闻领域缺乏一套规范性强、可用性高的成型知识图谱构建技术。针对上述两项研究困境，各大研究机构可与我社资源互补，真正提出一套面向新华社实际业务系统的知识图谱技术，相信对于解决新闻稿件文本智能分析问题上将发挥重要作用。

结论

本文介绍了在媒体融合发展的大趋势下，新闻媒体领域中文语义分析技术的智能化、知识化之路的研究与探索。

在智能化的中文语义分析技术部分，本文首先介绍了结合新闻稿件特性的智能化词语切分方面的研究，使分词效果更符合新闻媒体业务要求；其次，分别从应用场景出发，简要说明了语义分析算法，介绍了智能化主题分类、情感分类和自动摘要技术。

在知识化的中文语义分析技术部分，本文提出了结合新闻要素和特性的知识标签体系，并结合五类标签的实际特征，设计了基于语义紧密度挖掘与关键度排序的标签自动抽取算法；同时，面向新华社业务系统，对新闻媒体领域规范性强、可用性高的知识图谱技术进行了探索与展望。媒

参考文献

[1] 宗成庆 . 统计自然语言处理 [M]. 北京：清华大学出版社，2008.
[2] 李航 . 统计学习方法 [J]. 北京：清华大学出版社，2012.
[3] 俞士汶等 . 现代汉语语法信息词典详解 [M]. 北京：清华大学出版社，2003.

（作者单位：新华社技术局）